Characterization of the proteome of Theobroma cacao beans by nano-UHPLC-ESI MS/MS

Article

Accepted Version

Scollo, E., Neville, D., Oruna-Concha, M. J. ORCID: https://orcid.org/0000-0001-7916-1592, Trotin, M. and Cramer, R. ORCID: https://orcid.org/0000-0002-8037-2511 (2018) Characterization of the proteome of Theobroma cacao beans by nano-UHPLC-ESI MS/MS. Proteomics, 18 (3-4). 1700339. ISSN 1615-9853 doi: https://doi.org/10.1002/pmic.201700339 Available at https://centaur.reading.ac.uk/74661/

Dataset Brief

Characterization of the proteome of *Theobroma cacao* beans by nanoUHPLC-ESI MS/MS

Emanuele Scollo[1,2], David Neville[2], M. Jose Oruna-Concha[3], Martine Trotin[2] and Rainer Cramer[1*]

[1] Department of Chemistry, University of Reading, Reading RG6 6AD, UK

[2] Mondelēz International, Reading Science Centre, Reading RG6 6LA, UK

[3] Department of Food and Nutritional Sciences, University of Reading, Reading RG6 6AP, UK

*Address correspondence to:

Prof Rainer Cramer, Department of Chemistry, University of Reading, Whiteknights, Reading RG6 6AD, UK.

Tel.: +44-118-378-4550; e-mail: r.k.cramer@rdg.ac.uk

Running title: The proteome of *Theobroma cacao* beans

Nonstandard abbreviations: emPAI (exponentially modified protein abundance index); FDR (false discovery rate); BSA (bovine serum albumin); SPITC (4-sulfophenyl isothiocyanate);

Keywords: Theobroma cacao, cocoa beans, plant proteomics, storage proteins, cocoa bean proteome

Received: 09 08, 2017; Revised: 11 29, 2017; Accepted: 12 22, 2017

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201700339 .

Abstract

Cocoa seed storage proteins play an important role in flavour development as aroma precursors are formed from their degradation during fermentation. Major proteins in the beans of *Theobroma cacao* are the storage proteins belonging to the vicilin and albumin classes. Although both these classes of proteins have been extensively characterized, there is still limited information on the expression and abundance of other proteins present in cocoa beans. This work is the first attempt to characterize the whole cocoa bean proteome by nanoUHPLC-ESI MS/MS analysis using tryptic digests of cocoa bean protein extracts. The results of this analysis show that a total of 906 proteins could be identified using a species-specific *Theobroma cacao* database. The majority of the identified proteins were involved with metabolism and energy. Additionally, a significant number of the identified proteins were linked to protein synthesis and processing. Several proteins were also involved with plant response to stress conditions and defence. Albumin and vicilin storage proteins showed the highest intensity values among all detected proteins, although only seven entries were identified as storage proteins. A comparison of MS/MS data searches carried out against larger non-specific databases confirmed that using a species-specific database can increase the number of identified proteins, and at the same time reduce the number of false positives. The results of this work will be useful in developing tools which can allow the comparison of the proteomic profile of cocoa beans from different genotypes and geographic origins. Data are available via ProteomeXchange with identifier PXD005586.

The cocoa tree *Theobroma cacao* (family *Sterculiacae*) originates from the Amazon and Orinoco valleys and its natural habitats are in the tropical areas of South and

Central America. The name *Theobroma* is derived from the Greek words '*Theo*' (meaning god) and *'Broma'* (meaning food), referring to the Mayan and Aztec popular belief that chocolate was the food of the gods. Principal varieties of *Theobroma cacao* are Forastero, Criollo and Trinitario. Forastero varieties are regarded as "bulk cocoa in trade" and make up almost 95% of the cocoa's total worldwide production [1]. The cultivation of Criollo is limited to a few regions in Central America and Asia, as this population is susceptible to diseases and has a very low yield. The Trinitario type is native to Trinidad and includes all hybridization combinations of the Criollo and Forastero varieties. Both the Trinitario and the Criollo varieties produce the 'fine flavour' cocoa beans, which account for less than 5% of the total cocoa's world production [1].

Chocolate is basically made from processed cocoa beans. In order to produce the characteristic cocoa aroma, fermentation of cocoa beans is essential. During this process storage proteins in the cocoa beans are degraded by endogenous proteases with the release of peptides and amino acids, which are considered to be important flavour precursors for the generation of cocoa aroma during roasting. Under-fermented or not-fermented cocoa beans do not have the right amount of flavour precursors, and lack the cocoa aroma when roasted.

The major cocoa bean storage proteins are vicilin and albumin, representing 43% and 52% of the total cocoa seed proteins, respectively [2]. Other authors, however, have stated that vicilin represents only 23% of the soluble seed proteins, and albumin only 14.1% [3]. Polypeptides with molecular weights of 47 kDa, 31 kDa and 14.5 kDa are the predominant components of the vicilin fraction when analyzed by 1D-SDS-PAGE [2, 3]. The MALDI-TOF MS analysis of the tryptic digests of these

polypeptides confirmed that their amino acid sequences can be localized on a common precursor of ~66 kDa [4]. The main component of the albumin fraction is a polypeptide with a molecular weight of 21 kDa. This polypeptide is derived from a cDNA precursor that translates to give a 221-amino-acid polypeptide of 24 kDa [5]. The primary structure of the 21 kDa albumin protein has been characterized by LC-MS/MS analysis of its tryptic digests, confirming that the amino acid sequence of the mature expressed protein is nine amino acids shorter than the sequence expected from its encoding cDNA [6].

Recently, an analysis of the proteomic profile of *Theobroma cacao* pod husk was carried out by initial separation of intact proteins using 2D gel electrophoresis and subsequent *de novo* sequencing of SPITC-derivatized tryptic peptides of the excised gel bands using PSD-MALDI-TOF/TOF MS/MS [7]. The majority of the identified proteins could be linked to metabolism and energy, and a considerable proportion was involved with pod growth and development processes [7]. A similar procedure has been employed to perform proteomic analysis of *Theobroma cacao* embryos [8]. In this case, the majority of the identified proteins were related to genetic information processing, carbohydrate metabolism and stress response. MALDI-TOF MS was also employed to characterize the water-soluble portion of the proteomic seed extracts from different varieties of *Theobroma cacao* [9]. Most of the proteins detected with this approach showed molecular weights between 8-13 kDa, while a cluster at 21 kDa was attributed to albumin.

The study presented here is the first attempt of characterizing the whole cocoa bean proteome using a bottom-up shotgun approach, analyzing tryptic digests of cocoa

bean protein extracts by nanoUHPLC-ESI MS/MS. The reported results include a comparison of MS/MS data searches using different sequence databases.

Cocoa seeds were from ripe pods of the West African Amelonado variety harvested at the Cocoa Research Centre of the University of West Indies, St. Augustine, Trinidad. Approximately 240 beans from 6 different pods were combined together. The cocoa seeds were freeze-dried and ground. Two samples were then taken, one for a sample preparation workflow with fractionation and another for a workflow without fractionation. In both workflows, these samples were defatted using petroleum ether. Polyphenols were then extracted from the defatted cocoa seeds to avoid complex formation of polyphenols with proteins [2]. In the workflow with fractionation, albumin and vicilin fractions were extracted with a low and high ionic strength buffer, respectively. Proteins in the cocoa powder left after these buffer extractions were extracted with a solution containing chaotropic agents. This solution was also used for the protein extraction of the cocoa powder in the workflow without fractionation. An aliquot of each sample solution, containing a total amount of 10 µg of proteins based on the Bradford assay, was reduced and alkylated, and subsequently digested with trypsin. The tryptic digests were desalted and analyzed on a nanoUHPLC-MS/MS system consisting of an Orbitrap Fusion (Thermo Fisher Scientific, Hemel Hempstead, UK) mass spectrometer coupled to a Dionex Ultimate 3000 Nano RSLC (Dionex/Thermo Fisher Scientific).

For each extraction condition the MS/MS raw file was processed using Mascot Distiller software (Matrix Science Ltd, London, UK; Version 2.5.1.0). Mascot searches were carried out against the Cacao Matina1-6 Genome v1.1 *Theobroma cacao* database

(http://www.cacaogenomedb.org/Tcacao_genome_v1.1#tripal_analysis-downloads-box; accessed on 31st May 2015; 59,577 sequences; 23,720,084 residues), the NCBInr database (downloaded on 16th June 2015; 67,841,823 sequences; 24,324,060,020 residues), the Uniprot\Swissprot database (downloaded on 31st March 2014; 542,782 sequences; 193,019,802 residues), a custom NCBInr database with entries restricted to *Theobroma cacao* only (downloaded on 7th July 2015; 43,683 sequences; 19,146,837 residues) and a custom Uniprot\Tremble database with entries restricted to *Theobroma cacao* only (downloaded on 1st July 2015; 40,941 sequences; 17,501,566 residues).

Details of the sample extractions, LC-MS analysis and Mascot search parameters can be found in the Supporting Information.

The first results obtained in this study show that the recently published genome of *Theobroma cacao* [10] provides a relatively comprehensive protein database for proteomic analyses. Searches against this database of the MS/MS data obtained from tryptic digests returned a total of 906 and 704 proteins hits for the fractionated and unfractionated sample, respectively. About 86% of the proteins detected in the unfractionated sample (607 hits) were also detected in the fractions of the fractionated sample, confirming that the two samples had a high degree of overlap. As sample fractionation only increased the number of identified proteins by less than 30%, other methods leading to reduced analyte suppression effects typically observed with highly complex samples due to co-elution/analysis should be considered in the future. As for the fractionated sample, 590 protein identifications were recorded when the identifications of the water soluble and salt soluble fractions

were combined, of which 376 were also present in the urea soluble fraction (see Figure 1).

In order to assess whether searching different databases would yield a higher number of protein hits, searches using NCBInr and Uniprot\Swissprot databases with taxonomy *Viridiplantae*, and custom databases containing only *Theobroma cacao* entries from Uniprot\Tremble and NCBInr were also carried out. The Uniprot\Tremble version was chosen for the custom database as the Uniprot\Swissprot version returned only 7 hits (Endochitinase 1, CHI1_THECC; Arginine decarboxylase, CHI1_THECC; Casparian strip mebrane protein, SPE2_THECC; Vicilin, VCL_THECC; Maturase K, MATK_THECC; 21 kDa seed protein, ASP_THECC; Coat protein, COAT_CAYMV). The results of these searches are shown in Table 12.

The three *T. cacao*-specific databases led to the search results with the highest numbers of proteins being identified. The highest number of identified proteins (906) was obtained when searching the Cacao Matina1-6 Genome database. A slightly lower number of proteins were detected in the Uniprot/Tremble database (897) and the NCBInr database (870) when restricted to *Theobroma cacao* entries. Only 364 proteins could be identified from a search using the Uniprot\Swissprot database with the taxonomy set to *Viridiplantae*, while 759 proteins were identified when searching the same data against the NCBInr database with taxonomy *Viridiplantae*. These results confirm previous plant proteomic results [11] and show that the use of a more species-specific database can increase the number of identified proteins.

Proteins were also classified according to their main biological function using the results from the search of the Cacao Matina1-6 Genome database of the fractionated sample (906 protein identifications). A graphical overview of the protein

classification is shown in Figure 2 with two representations, providing the abundance-weighted and unweighted percentage for each protein class (function group). The percentage of each protein function for the abundance-weighted classification was calculated by summing the normalised responses of proteins detected in all three fractions.

A complete list of all identified proteins can be found in the Supporting Information Table S1.

The majority of the identified proteins were linked to the function group *metabolism and energy* with 505 entries which accounted for 55.7% of the total number of identified proteins. The most abundant proteins within this function group belonged to the oleosins family. There is a direct correlation between the content of oil in seeds and the level of oleosin [12]. However, the exact role of oleosins in oil accumulation has not yet been elucidated [12]. Several phosphoglycerate kinases and glyceraldehyde-3-phosphate dehydrogenases, two enzyme families which play an important role in glycolysis and the Calvin cycle, were among the most abundant proteins linked to *metabolism and energy*. Proteins belonging to the aldolases family were also present at significant levels among the proteins involved with *metabolism and energy*.

Proteins involved with *protein synthesis and processing* accounted for 25.7% of the total number of identified proteins (233 entries). The main proteins within this group were molecular chaperones. Several ribosomal proteins were also detected within this category. A significant number of identified proteins (44 entries and 4.9% of the total protein number) could be linked to *defence and stress*. Several heat-shock proteins and chitinase were also found within this function group. Heat-shock

proteins are synthesized when the plant is exposed to adverse environmental factors and therefore determine the ability of plants to survive under such unfavourable conditions [13]. Chitinases play a crucial role in plant defence against pathogens as they hydrolyze chitin, which is a structural component of the cell wall of many phytopathogenic fungi [14].

Vicilin and albumin were classified as *storage proteins* and showed the highest BSA-normalized emPAI value among all detected proteins, although only seven entries could be linked to this function group. However, if the relative percentages of the classified proteins are based on the sum of the BSA-normalized emPAI values of all proteins within the same function group, *storage proteins* are the third most abundant protein group after *metabolism and energy* and *protein synthesis and processing* (see Figure 2, lower pie chart). The ratio albumin/vicilin in the water soluble and salt soluble fractions were 3.5 and 0.9, respectively. The relative amounts of vicilin and albumin compared to the total protein amount in the whole fractionated sample were 3.9% and 11.5%, respectively. These levels are lower than the reported values of between 43% and 23% for vicilin [2, 3], and 52% and 14% for albumin [2, 3]. The higher values in the literature are possibly due to the inefficiency of selective solubilisation of proteins [2] and differences in the detection and separation of proteins analysed by SDS-PAGE [3], which as a result can lead to an overestimation of protein amount.

In conclusion, this first attempt to characterize the whole cocoa bean proteome by nanoUHPLC-ESI MS/MS analysis of tryptic digests of cocoa bean protein extracts led to a total of more than 900 protein identification. The presented methodology may benefit from further optimization in terms of protein extraction and

chromatographic separation. However, its current performance and the dataset obtained already provide a good platform for studies aimed at gaining a better understanding of the proteomic profile of cocoa beans and present the largest proteome dataset for cocoa beans to date.

Data supporting the results of this dataset brief are available in the PRIDE (Proteomics Identifications Database) partner repository at the European Bioinformatics Institute, PXD005586 (http://www.ebi.ac.uk/pride/).

Acknowledgement

This work was supported by the BBSRC (grant BB/M012387/1) through access to the Orbitrap Fusion instrument. The authors are grateful to the University of West Indies Cocoa Research Centre for providing cocoa beans.

References

[1] Lima, L. J. R., Almeida, M. H., Nout, M. J. R., Zwietering, M. H., *Theobroma cacao L*., "The Food of the Gods": Quality Determinants of Commercial Cocoa Beans, with Particular Reference to the Impact of Fermentation. *Critical Reviews in Food Science and Nutrition* 2011, *51*, 731-761.

[2] Voigt, J., Biehl, B., Wazir, S. K. S., The Major Seed Proteins of *Theobroma cacao L. Food Chemistry* 1993, *47*, 145-151.

[3] Lerceteau, E., Rogers, J., Petiard, V., Crouzillat, D., Evolution of cacao bean proteins during fermentation: a study by two-dimensional electrophoresis. *Journal of the Science of Food and Agriculture* 1999, *79*, 619-625.

[4] Kratzer, U., Frank, R., Kalbacher, H., Biehl, B., Woestemeyer, J., Voigt, J., *Food Chemistry* 2009, pp. 903-913.

[5] Spencer, M. E., Hodge, R., Cloning and sequencing of the cDNA-encoding the major albumin of *Theobroma cacao*-Identification of the protein as a member of the Kunitz protease inhibitor family. *Planta* 1991, *183*, 528-535.

[6] Kochhar, S., Gartenmann, K., Juillerat, M. A., Primary structure of the abundant seed albumin of *Theobroma cacao* by mass spectrometry. *Journal of Agricultural and Food Chemistry* 2000, *48*, 5593-5599.

[7] Awang, A., Karim, R., Mitsui, T., Proteomic analysis of *Theobroma cacao* pod husk. *Journal of Applied Glycoscience* 2010, *57*, 245-264.

[8] Noah, A. M., Niemenak, N., Sunderhaus, S., Haase, C., Omokolo, D. N., Winkelmann, T., Braun, H.-P., Comparative proteomic analysis of early somatic and zygotic embryogenesis in *Theobroma cacao* L. *Journal of Proteomics* 2013, *78*, 123-133.

[9] Bertazzo, A., Agnolin, F., Comai, S., Zancato, M., Costa, C. V. L., Seraglia, R., Traldi, P., The protein profile of *Theobroma cacao* L. seeds as obtained by matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Communications in Mass Spectrometry* 2011, *25*, 2035-2042.

[10] Motamayor, J. C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone, D., III, Cornejo, O., Findley, S. D., Zheng, P., Utro, F., Royaert, S., Saski, C., Jenkins, J., Podicheti, R., Zhao, M., Scheffler, B. E., Stack, J. C., Feltus, F. A., Mustiga, G. M., Amores, F., Phillips, W., Marelli, J. P., May, G. D., Shapiro, H., Ma, J., Bustamante, C. D., Schnell, R. J., Main, D., Gilbert, D., Parida, L., Kuhn, D. N., The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biology* 2013, *14*, 1-48.

[11] Bryant, L., Flatley, B., Patole, C., Brown, G. D., Cramer, R., Proteomic analysis of Artemisia annua - towards elucidating the biosynthetic pathways of the antimalarial pro-drug artemisinin. *BMC Plant Biology* 2015, *15*.

[12] Parthibane, V., Rajakumari, S., Venkateshwari, V., Iyappan, R., Rajasekharan, R., Oleosin Is Bifunctional Enzyme That Has Both Monoacylglycerol Acyltransferase and Phospholipase Activities. *Journal of Biological Chemistry* 2012, *287*, 1946-1954.

[13] Al-Whaibi, M. H., Plant heat-shock proteins: A mini review. *Journal of King Saud University Science* 2011, *23*, 139-150.

[14] Punja, Z. K., Zhang, Y. Y., PLANT CHITINASES AND THEIR ROLES IN RESISTANCE TO FUNGAL DISEASES. *Journal of Nematology* 1993, *25*, 526-540.

Table 1. Number of identified proteins by nanoUHPLC-ESI MS/MS at 1% FDR for data searches using the search engine Mascot and different protein sequence databases

Database	No. of Sequences in Database	No. of Identified Proteins
Cacao Matina1-6 Genome (*T. cacao)*	59,577	906
Uniprot\Swissprot (*Viridiplantae*)	34,907	364
NCBInr (*Viridiplantae*)	3,047,619	759
Uniprot\Tremble (*T. cacao*)	40,941	897
NCBInr (*T. cacao*)	43,683	870

Figure Legend

Figure 1. Venn diagram showing the numbers of proteins identified by nanoUHPLC-ESI MS/MS from a fractionated *T. cacao* sample searched against the Cacao Matina 1-6 Genome database published by Motamajor *et al.* (2013). The protein fractions are labelled as follows: WS, water soluble fraction; SS, salt soluble fraction; US, urea soluble fraction.

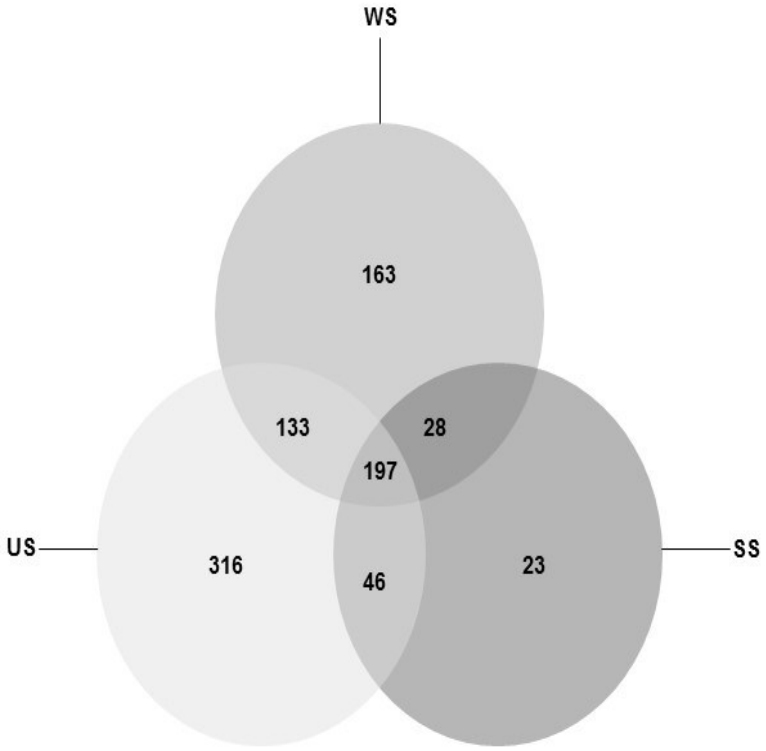

Figure 2. Classification of cocoa bean proteins based on their function. The percentages in the upper pie chart represent the number of proteins in each function group relative to the total number of proteins. Each function group is also labelled with the number of proteins. The lower pie chart provides the sums of the BSA-normalized emPAI values of the proteins in each function group relative to the total sum of the BSA-normalized emPAI values of all proteins. The function group labels are as follows: ME, metabolism and energy; PSP, protein synthesis and processing; SP, storage proteins; MT, membrane transport; ST, signal transduction; UN, unclassified; CS, cell structure; DNA, DNA synthesis and processing; GD, growth and development; DFS, defence and stress.